

ノード属性を考慮したグラフのまとめ発見手法

寺田 愛花 (指導教員：瀬々 潤)

1 はじめに

近年、遺伝子ネットワーク、Web、ソーシャルネットワークなど、様々なデータがグラフで表現されている。しかしこのようなグラフは、頂点数、辺の本数が膨大であり、単に頂点と辺を与えたグラフから情報を読み取することは困難である。それゆえ、グラフのまとめを発見する技術が必要とされている。

まとめの発見手法としてクラスタリングがあり、辺が高密度に張られている頂点集合を発見する手法 [1] か疎な部分を分断する方法が一般的であるこの手法では、関係の強い頂点集合を求めることはできるが、グラフ全体の構造を見ることは難しい場合がある。

大域的な構造を見る手法として k -SNAP [2] がある。 k -SNAP は、ラベル付きグラフから、クラスタ内の頂点のラベルが等しく、頂点の隣接するクラスタが同一となるように分割する。しかし k -SNAP は、同一クラスタ内の頂点はラベルが等しいという制約から、データ内のラベル数が k 以下でないと解析できないという制約がある。

そこで本研究では、グラフの隣接構造を用いた頂点の併合と、ラベル情報を用いた頂点の分割を行うことで、 k 種類以上のラベルを持つデータに対してもグラフ全体の構造を見ることができるアルゴリズムを提案する。

また頂点に遺伝子、辺に遺伝子ネットワーク、ラベルに遺伝子の機能を与えたネットワークを解析し、細胞内のネットワークの構造を抽出する。

2 提案手法

提案手法は頂点の併合と分割の2段階で構成されるアルゴリズムである。

1. ラベルの値が等しい頂点でクラスタを構成
2. $k/2$ 個まで類似するクラスタを併合
3. クラスタの個数が k 個になるまでクラスタ間の強い関係を抽出する分割をする

グラフ $G = (V, E)$ とする。 V は頂点集合、 E は辺の集合である。頂点にはラベルを与え、このラベル集合を $A = \{a_1, a_2, \dots, a_t\}$ と表す。頂点 v がラベル a_i を有する場合 $a_i(v) = 1$ 、それ以外の場合 0 とする。またグラフ G の分割を $\Phi = \{G_1, G_2, \dots, G_k\}$ と表記する。頂点 v は必ず一つのクラスタ G_i に属し、 $G_i \neq \phi$ ($\forall G_i \in \Phi$) かつ $G_i \cap G_j = \phi$ ($i \neq j$) である。頂点 v の属するクラスタを $\Phi(v)$ とする。

2.1 ラベルによるクラスタの構成

まず頂点をラベルの値で分割する。そのために $u, v \in V$ に対し、 $\forall a_i \in A$ で $a_i(u) = a_i(v)$ ならば $\Phi(u) = \Phi(v)$ 、それ以外の場合、 $\Phi(u) \neq \Phi(v)$ となるようにクラスタを構成する。 t 個のクラスタが形成される。

2.2 クラスタの併合

併合するクラスタは、クラスタ間の辺の伸び方が類似するクラスタである。この類似性により、クラスタ

間距離 $D(G_i, G_j)$ を定義する。

クラスタ間の辺の伸びの類似性はコサイン距離で計測する。 $|E(G_i, G_j)|$ を G_i と G_j の間にある辺の本数とし、辺の割合 r_{ij} を定義する。

$$r_{ij} = \begin{cases} \frac{2|E(G_i, G_j)|}{|G_i|(|G_i|-1)} & (i = j) \\ \frac{|E(G_i, G_j)|}{|G_i||G_j|} & (i \neq j) \end{cases} \quad (1)$$

G_i から G_j へ伸びる辺の割合 r_{ij} を各要素としたベクトルを $\mathbf{r}_i = (r_{i1}, r_{i2}, \dots, r_{in})$ とする。辺の伸び方が類似しているクラスタは、このベクトルの向きが類似している。よって、本研究では、コサイン距離を用いてクラスタ間の類似度を計測する。

$$D(G_i, G_j) = 1 - \frac{\mathbf{r}_i \cdot \mathbf{r}_j}{|\mathbf{r}_i||\mathbf{r}_j|} \quad (2)$$

この距離を全クラスタ間で算出し、値が最小のクラスタ G_i と G_j を併合する。この操作を繰り返し、クラスタ数が $k/2$ 個になるまで併合する。

2.3 クラスタ間の関係を抽出する分割

クラスタを併合した後、クラスタ間の関係を保ちつつ分割する。この手順は k -SNAP 法と同等である。

G_i に属する頂点で、 G_j と辺を持つ頂点集合を $P_{G_j}(G_i) = \{u | u \in G_i, \exists v \in G_j, (u, v) \in E\}$ と定義する。クラスタ G_i が G_j に対して関係が強いとは、クラスタ内にこのような頂点が多いことである。そこで、クラスタ G_i と G_j の強さの割合を次式で定義する。

$$p_{i,j} = \frac{|P_{G_j}(G_i)| + |P_{G_i}(G_j)|}{|G_i| + |G_j|} \quad (3)$$

我々が抽出したいクラスタは $p_{i,j}$ が 0 または 1 に近いクラスタである。そのため、クラスタ G_i の G_j に対する分割の良さを以下のように定義する。

$$\delta_{G_j}(G_i) = \begin{cases} |G_i| - |P_{G_j}(G_i)| & (p_{i,j} > 0.5) \\ |P_{G_j}(G_i)| & (\text{上記以外}) \end{cases} \quad (4)$$

この指標は値が小さい方が良い分割である。そのため、この値が最大のクラスタを分割する。

クラスタ G_i の指標を $CT(G_i) = \max_{G_j} \{\delta_{G_j}(G_i)\}$ とする。この値が最大クラスタを G_s とし、 $G_t = \arg \max_{G_j} \{\delta_{G_j}(G_s)\}$ であるクラスタ G_t に対して辺を持つ頂点集合と持たない頂点集合に分割することで、クラスタ間の強い関係を抽出していく。これを、クラスタ数が k 個になるまで繰り返す。

3 遺伝子データでの実行

遺伝子ネットワークを解析し、機能間の関係を抽出する。解析するグラフは、頂点に遺伝子、辺に遺伝子ネットワーク、ラベルに遺伝子の機能を与える。その頂点数は 1,131 個、辺の本数は 44,209 本、ラベルは 103 個である。

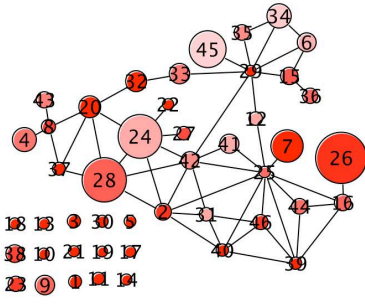


図 1: 本研究の分割結果

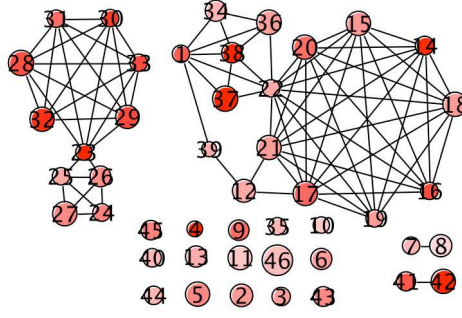


図 2: 一般的なグラフクラスタリングの結果

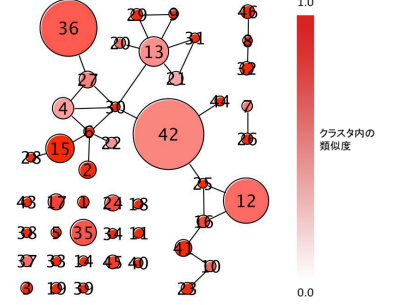


図 3: 併合のみ行った結果

このネットワークを k -SNAP で解析すると、ラベルの制約から 103 個のクラスタが構成されるため、46 個に分類することはできない。

そこで本研究のアルゴリズムと比較するため、頻繁に用いられるグラフクラスタリング手法 [3] と、ラベルによるクラスタの構成後、46 個まで併合のみ行う手法による解析も行った。これらの結果を、クラスタ内の頂点の類似性と、クラスタ間の関係の強さから比較する。

クラスタリングの結果を、 G_i を一つの頂点 i とし、多くの頂点が属するクラスタほど大きい頂点を与えたグラフで表す。

クラスタ G_i 内の類似性を、属する頂点のラベルの類似度として次式で定義する。

$$ap(G_i) = \frac{\max_{a_j \in A} \{ \sum_{v \in G_i} a_j(v) \}}{|G_i|} \quad (5)$$

結果のグラフでは、頂点の色が赤いほどクラスタ内の類似度が高く、白いほど低いことを表している。

また、クラスタ間の強い関係を見るために、クラスタ G_i と G_j 間の関係の強さを次式で定義する。

$$ep(G_i, G_j) = \max \left\{ \frac{|P_{G_j}(G_i)|}{|G_i|}, \frac{|P_{G_i}(G_j)|}{|G_j|} \right\} \quad (6)$$

結果のグラフには、 $ep(G_i, G_j) > 0.7$ のクラスタ間に辺を与えた。

本手法の結果が図 1、一般的なグラフクラスタリングによる結果が図 2、ラベルの値による分類を行った後、併合のみ行った結果が図 3 である。

3.1 クラスタ内の類似性の比較

$ap(G_i)$ の値が大きいほどクラスタ内の頂点は類似している。そこで $\frac{|\{G_i | ap(G_i) > 0.5, \forall G_i \in \Phi\}|}{k}$ を結果の比較のために算出した。その割合は、本研究の結果が約 0.84、一般的なグラフクラスタリングの結果が 0.52、ラベルの分類後、併合のみ行った結果が 0.93 であった。

本研究の結果と、併合のみを行った結果のクラスタ内の類似性が高いことから、辺の伸び方が類似しているクラスタを併合することで、類似性の高い頂点を同一クラスタに分類できることがわかる。一方で、一般的な手法では、クラスタ内の類似性はそれほど高くなかった。

3.2 クラスタ間の関係の比較

実行結果を、クラスタ間の関係の強さから比較する。 $\delta_{G_j}(G_i)$ は、クラスタ G_i の G_j に対する分割が良い方

が小さい値をとる。そこで、分割 $\Phi = \{G_1, G_2, \dots, G_k\}$ の指標を次式で定義する。

$$\Delta(\Phi) = \frac{\sum_{G_i, G_j \in \Phi} (\delta_{G_j}(G_i) + \delta_{G_i}(G_j))}{k} \quad (7)$$

この値を算出すると、本手法は 15.84、一般的なクラスタリングの結果が 24.76、併合のみ行った結果が 21.06 である。

この値は小さい方が良い分割であるため、本研究のアルゴリズムが最もクラスタ間の密な関係が抽出できることがわかる。

3.3 結果の考察

遺伝子ネットワークで、本研究のアルゴリズムと、一般的なグラフクラスタリング手法、ラベルによる分類後、併合のみ行う三種類の手法を比較した。一般的な手法では、クラスタ内の類似性が低く、クラスタ間の関係も強くなかった。併合のみを行った手法では、クラスタ内の類似性は高いが、これもクラスタ間の関係を見ることはできなかった。本手法では、クラスタ内の類似性が高く、またクラスタ間の強い関係を最も抽出できていた。

4 今後の課題

本手法では、クラスタの併合の終了条件を $k/2$ 個としたが、この終了条件を分割の指標を使って求めるように改良したい。 $\Delta(\Phi)$ ではクラスタ数が異なる場合に適切な比較ができないため、新たな指標を定義する必要がある。

また、この指標を用いて併合と分割を繰り返し行い、より良いクラスタリングの結果を求めるように拡張することもできるだろう。

参考文献

- [1] 林幸雄. ネットワーク科学の道具箱. 近代科学社, 2007.
- [2] Yuanyuan Tian, Richard A. Hankins, and Jignesh M. Patel. Efficient aggregation for graph summarization. In *SIGMOD*, 2008.
- [3] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J.SCI.COMPUT.*, 1998.